ITC 14 / J. Labetoulle and J.W. Roberts (Editors)

PUBLICATION DATE: 06.06.94
(further bibliographic data on next page)

p.401–410 = (10)

Tariffs and effective bandwidths in multiservice networks

F.P. Kelly

Statistical Laboratory, University of Cambridge, 16 Mill Lane, Cambridge CB2 1SB, England, UK

We describe a framework which allows the investigation of possible tariffs for high speed multiservice networks, and discuss a simple tariff that encourages the cooperative sharing of information and effort between users and the network. The tariff has the property that if a user optimally selects the tariff to apply to a call at the time of the call's admission then the network can estimate the effective bandwidth of the call. A further property becomes apparent if a user can, with some effort, improve its prediction of the statistical properties of a call: the incentive for improved prediction provided by the tariff precisely matches the consequent expected reduction in the effective bandwidth required from the network. In the case of on/off sources with a policed peak rate the tariff takes a very simple form: a charge $a(m)$ per unit time and a charge $b(m)$ per unit of traffic carried, where the pair $(a(m), b(m))$ are fixed by a declaration m, made by the user at the time of call admission, of the expected rate of the source.

1. INTRODUCTION

The trend of current developments in telecommunication networks is towards systems which will allow a number of widely disparate traffic streams to share the same broadband channel ([2],[12],[13]). A call, which might be a mixture of voice, video and data, would appear to the network as a stream of cells, and the hope is that calls with a broad range of burstiness characteristics can be efficiently integrated, through statistical multiplexing, to share a common resource. In recent years a number of papers ([3],[4],[5],[6],[7],[8],[15]) have provided some basis for this hope, by showing that it is possible to associate an effective bandwidth with a source type such that, provided the sum of the effective bandwidths of the sources using a resource is less than a certain level, then the resource can deliver a performance guarantee, expressed in terms of the probability that delay exceeds a threshold or that a cell is lost. The effective bandwidth depends on characteristics of the source such as its mean and peak rate and there is by now a large literature on methods for policing the peak and mean rates of a call. The underlying idea is that at call admission a contract would be made between user and network specifying in more or less detail the statistical properties of the call, and

that policing mechanisms would enforce the contract. For a critical review of work on policing see Roberts [13]: policing peak rates appears to be relatively easy (for example, by use of a leaky bucket or by enforcing a minimum gap between successive cells of a stream [1]), but there is considerable evidence that mean rate policing is impractical.

The need for networks to operate in a public (and therefore potentially non-cooperative) environment has encouraged the notion that calls must be effectively policed; unfortunately policing may limit many of the advantages of a high speed multi-service network, such as the network's inherent flexibility to deal with a call composed of varying and uncertain mixtures of voice, video and data. What is needed is a mechanism that trades off the user's uncertainty about a commencing call against the network's requirement to statistically multiplex calls in an efficient manner. In this paper we describe a model for tariffs within which this issue may be explored, and we use the model to develop a simple tariff structure which encourages the cooperative sharing of information and effort between users and the network.

The organization of the paper is as follows. In Section 2 we review the concept of effective bandwidth, originally due to Hui [7]. In Section 3 we consider the case of on/off sources of known peak rate, but where the mean rate of a source may not be known with certainty, even to the user responsible for the source. We show that the effective bandwidth of a source depends on just the user's *expected* mean rate, and we describe a tariff structure which encourages the user to accurately declare this quantity. A further property of the tariff structure becomes apparent if a user can, with some effort, improve its predictions of the statistical properties of a call: the incentive for improved prediction provided by the tariff structure precisely matches the consequent expected reduction in the effective bandwidth required from the network. The tariff structure has a very simple form: a charge $a(m)$ per unit time and a charge $b(m)$ per unit of traffic carried, where the pair $(a(m), b(m))$ are fixed by a declaration m, made by the user at the time of call admission, of the expected mean rate of the source. In Section 4 we provide a numerical example of the tariff, observe its dependence on the extent of statistical multiplexing possible, and discuss the balance the tariff implies between charges per unit time and charges per unit of traffic carried. In Section 5 we note certain generalizations of our results [10]: all that is necessary for our approach to carry through is that the effective bandwidth of a source be expressed as a concave function of the expectation of a measurable quantity.

2. EFFECTIVE BANDWIDTHS

Suppose that J sources share a single resource of capacity C, and let X_j be the load produced by source j. Assume that $X_j, j = 1, 2, \ldots, J$, are independent random variables, and let the distribution function of X_j be F_j. Can the resource cope with the superposition of the J sources? More precisely, can we impose a condition on $F_1, F_2, \ldots, F_J$ which ensures that

$$P\{\sum_{j=1}^{J} X_j > C\} \leq e^{-\gamma} \tag{1}$$

for a given value of γ? The answer to this question is, by now, fairly well understood. There are constants α, K (depending on γ and C) such that if

$$\sum_{j=1}^{J} B(F_j) \leq K\,, \tag{2}$$

where

$$B(F_j) = \frac{1}{\alpha} \log \mathrm{E} e^{\alpha X_j} \tag{3}$$

then condition (1) is satisfied. The expression (3) is called the *effective bandwidth* of source j. This result, originally due to Hui ([7],[8]), has been generalized in a variety of directions. For example, in [9] it is shown that if the resource has a buffer, and if the load produced by source j in successive time periods is a sequence of independent bursts each with distribution F_j, then the probability the delay at the resource exceeds b time periods will be held below $e^{-\gamma}$ provided inequality (2) is satisfied, with B again given by equation (3), where $K = C$ and $\alpha = \gamma/(bC)$. It is by now known that for quite general models of sources and resources it is possible to associate an effective bandwidth with each source such that, provided the sum of the effective bandwidths of the sources using a resource is less than a certain level, then the resource can deliver a performance guarantee (see [4],[5],[15]); of course, for more general models the definition of the effective bandwidth is necessarily more complicated than the simple form (3). It is still not clear when the more general definitions of effective bandwidth will be needed: many simple and effective methods of congestion control render the simple form (3) the appropriate definition (see, for example, [14]).

In this paper we shall concentrate mainly on the simple form (3), leaving a brief mention of more general definitions of effective bandwidth to Section 5. Indeed we shall simplify further, to the case of an on/off source of peak rate h and mean rate M for which

$$P\{X = 0\} = 1 - \frac{M}{h} \qquad P\{X = h\} = \frac{M}{h}\,.$$

The effective bandwidth (3) of such a source is then

$$B(M) = \frac{1}{\alpha} \log \left[1 + \frac{M}{h}\left(e^{\alpha h} - 1\right)\right]\,. \tag{4}$$

3. SOURCES OF KNOWN PEAK RATE

Suppose that a source is an on/off source of known peak rate h cells per unit time,

but that its mean rate may not be known with certainty, even to the user responsible for a source. Suppose that a user, about to make a call, has a prior distribution G for the mean rate M of the call. The user might, for example construct the distribution G by recording the observed mean rates on past calls. Then the expected mean rate of the call is

$$\mathsf{E}_G M = \int_0^h x dG(x).$$

If the network knew the prior distribution G for the mean rate M, then the network would determine the effective bandwidth of the call, from equations (3) and (4), as

$$\frac{1}{\alpha}\log \mathsf{E}e^{\alpha X} = \frac{1}{\alpha}\log \mathsf{E}_G \mathsf{E}(e^{\alpha X}|M) = \frac{1}{\alpha}\log \mathsf{E}_G\left[1+\frac{M}{h}(e^{\alpha h}-1)\right]$$
$$= \frac{1}{\alpha}\log\left[1+\frac{\mathsf{E}_G M}{h}(e^{\alpha h}-1)\right]. \tag{5}$$

But expression (5) is just the effective bandwidth if M is not random, but identical to its mean value under G. We see that since the source is on/off with known peak rate the network need only know $\mathsf{E}_G M$, the user's expected mean rate; further detail about the distribution G does not influence the effective bandwidth, and would be superfluous for the network to even request.

How, then, should the network encourage the user to assess and to declare the user's expected mean rate? We next investigate whether the tariff structure might be used to provide the appropriate amount of encouragement.

Suppose that, before a call's admission, the network requires the user to announce a value m, and then charges for the call an amount $f(m;M)$ per unit time, where M is the measured mean rate for the call. We suppose that the user attempts to select m so as to minimize $\mathsf{E}_G f(m;M)$, the expected cost per unit time: call a minimizing choice of m, $\hat{m}$ say, an *optimal* declaration for the user. What properties would the network like the optimal declaration $\hat{m}$ to have? Well, first of all the network would like to be able to deduce from $\hat{m}$ the user's expected mean rate $\mathsf{E}_G M$, and hence the effective bandwidth (5) of the call. A second desirable property would be that the expected cost per unit time under the optimal declaration $\hat{m}$ be proportional to the effective bandwidth of the call (or, equivalently, *equal* to the effective bandwidth under a choice of units). In [10] it is shown that these two requirements essentially characterize the tariff $f(m;M)$ as

$$\hat{f}(m;M) = a(m) + b(m)M, \tag{6}$$

defined as the tangent to the curve $B(M)$ at the point $M = m$ (see Figure 1). The key property used in the proof [10] is the strict concavity of $B(M)$ as a function of M. By a simple differentiation, the coefficients in expression (6) are given by

$$b(m) = \frac{e^{\alpha h}-1}{\alpha[h+m(e^{\alpha h}-1)]}, \qquad a(m) = B(m) - mb(m).$$

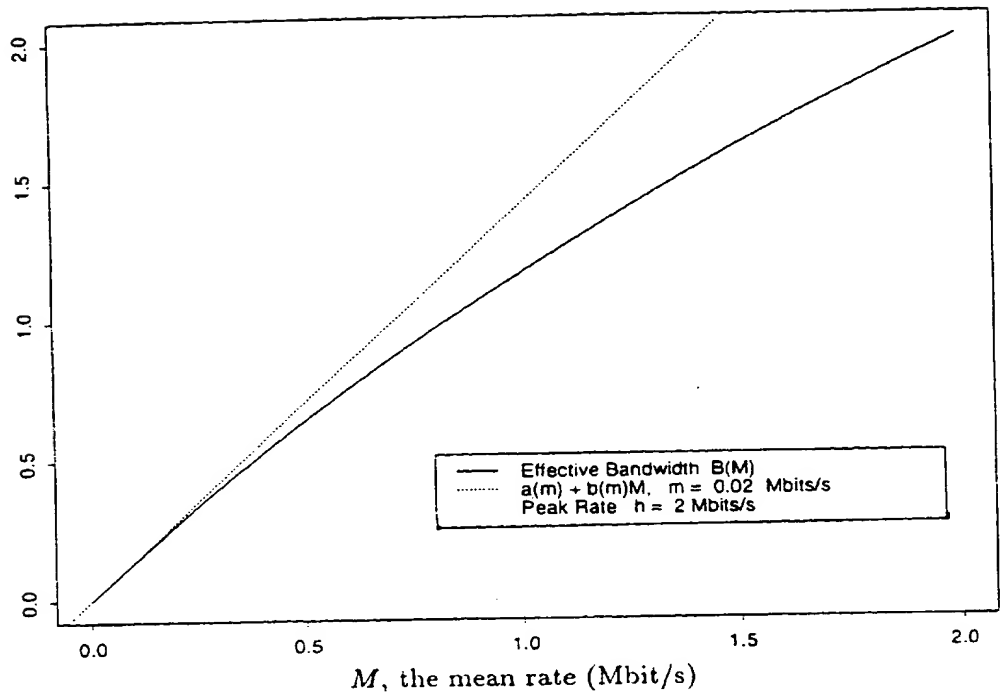


Figure 1. Effective bandwidth as a function of mean rate.

The properties characterizing the tariff $\hat{f}$ have many interesting and desirable consequences. For example, suppose that a user can, with some effort, improve its prediction of the statistical properties of a call. A crude method of deciding upon the declaration m might be to take the average of the measured means for the last n calls, but more sophisticated methods are possible: if the user is an organization containing many individuals, the user might observe the identity of the individual making the call, the applications active on that individual's desktop computer, as well as the called party, and utilize elaborate regression aids to make the prediction m. Is it worth the effort? Formally, let us suppose the user may construct a proxy indicator Y which can be observed by the user before call admission and which may be related to some greater or lesser extent with the future mean rate M of the call. The expected cost per unit time if the user observes the indicator Y and then optimizes the choice of tariff is

$$\mathsf{E}_G(\hat{f}(\mathsf{E}_G(M|Y), M)|Y) = \hat{f}(\mathsf{E}_G(M|Y), \mathsf{E}_G(M|Y)) = B(\mathsf{E}_G(M|Y)),$$

where G is now the user's joint prior distribution for Y and M. Thus the construction and utilization of the proxy indicator Y reduces the expected cost per unit time of the call from $B(\mathsf{E}_G M)$ to $\mathsf{E}_G(B(\mathsf{E}_G(M|Y)))$. But this is precisely the expected reduction in the effective bandwidth required from the network. This is an important property: users should not be expected to do more work determining the statistical properties of their calls than is justified by the benefit to the network of better characterization.

The above property is, of course, just a rephrasing of the property that the expected cost per unit time under the optimal declaration is equal to the effective bandwidth. If we do not insist on this property, but require only that the optimal declaration for the user be the mean $\hat{m} = \mathsf{E}_G M$, then many tariff structures are possible. For example, for any differentiable strictly concave function $F(M)$, let

$$f(m; M) = a'(m) + b'(m)M$$

be the tangent to the curve $F(M)$ at $M = m$. Then the optimal strategy for the user under the tariff structure f is to declare the mean $\hat{m} = \mathsf{E}_G M$. Tariffs that are non-linear in M can be designed to encourage declaration of, for example, percentiles of the distribution G: if F is any continuous strictly increasing function then the optimal declaration under the tariff

$$f(m; M) = F(m) + \gamma[F(M) - F(m)]^+$$

is the $100(1 - \gamma^{-1})$ percentile of the distribution G. The case $\gamma^{-1} = \frac{1}{2}$ encourages declaration of the median, which may approximate the mean. A tariff of a related form is discussed in [11] and in [13], pp. 65–6.

We end this section with a brief comment on the independence assumption underlying our approach. The independence of the loads produced by different sources is an essential feature of the analysis leading to the linear constraint (2), and the concept of an effective bandwidth. Within the framework of this section, this corresponds to the assumption that the errors made by different users in predicting their own mean rates are independent, in addition to the assumption that, conditional on mean rates, sources are on or off independently. While we judge both assumptions to be fairly plausible, one can certainly devise scenarios where either or both fail.

Table 1

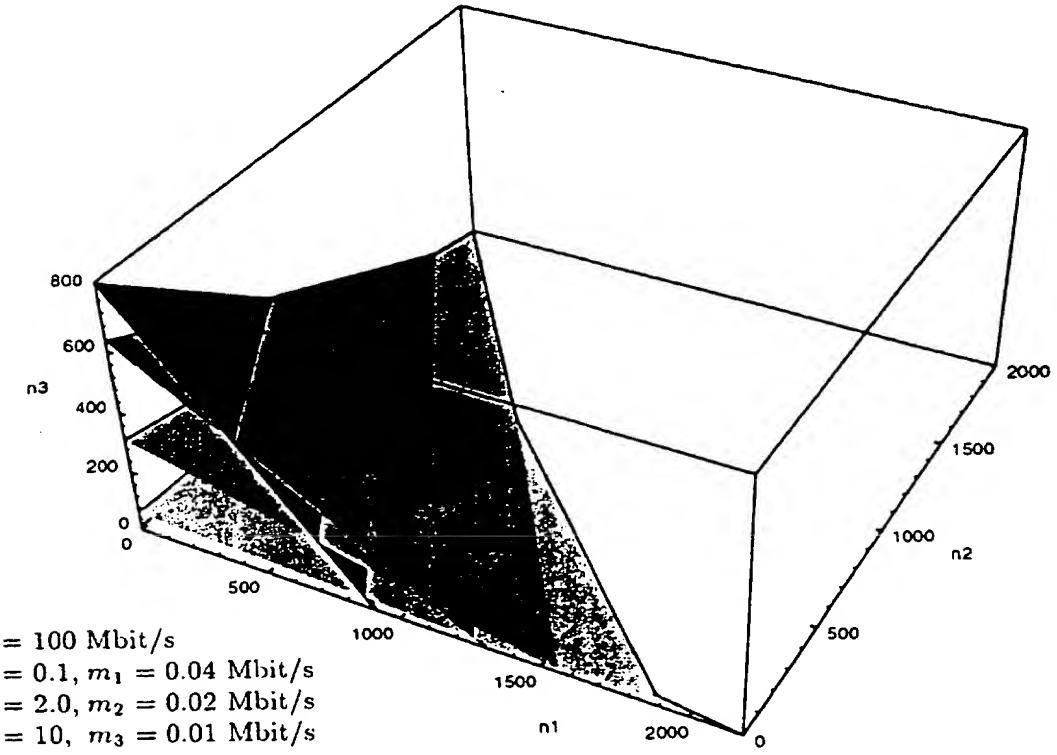
Service type	Rate (Mbits/s)		Charge	
	Peak	Mean	variable (s^{-1})	fixed (Mbit^{-1})
1	0.1	0.04	2.7×10^{-4}	1.0
2	2.0	0.02	1.3×10^{-4}	1.4
3	10.0	0.01	1.1×10^{-3}	7.9
	h	M	$a(M)$	$b(M)$

4. A NUMERICAL EXAMPLE*

We now illustrate the preceding sections with a numerical example. Suppose that the predominant traffic offered to a link of capacity 100 Mbit/s falls into three categories, with peak and mean rates as described in Table 1. Let $n1$, $n2$ and $n3$ be the numbers of the respective source types that have been admitted. Then, provided $(n1, n2, n3)$ lies

* I am grateful to James Tebboth for producing Figure 2 and to him and Stephen Turner for help with the other Figures of this paper.

towards the origin of any one of the planes illustrated in Figure 2, the probability of resource overload will be less than exp(−16): see [6],[7],[8],[9]. Let us select the second plane in Figure 2, intersecting the $n3$ axis at about $n3 = 625$: this corresponds to the choice $\alpha = 0.333$ in expression (3).



$C = 100$ Mbit/s
$h_1 = 0.1, m_1 = 0.04$ Mbit/s
$h_2 = 2.0, m_2 = 0.02$ Mbit/s
$h_3 = 10, \ m_3 = 0.01$ Mbit/s

Figure 2. Acceptance region.

While the predominant traffic may be of types 1, 2 and 3, the network operator may not want to constrain traffic to these three types, and, in particular, may be prepared to accept traffic of known peak rate with just a declaration of the expected mean rate. What tariff should be chosen for, say, traffic with peak rate 2 Mbit/s? In Figure 1 the effective bandwidth for such traffic is shown as a function of the mean rate, together with the tangent to this curve corresponding to the selection $m = 0.02$ Mbit/s. The coefficients $a(m)$ and $b(m)$ of the tariff (6) are shown in Figure 3.

Table 2

Service	Rate (Mbits/s)		Charge	
type	Peak	Mean	variable (s^{-1})	fixed (Mbit^{-1})
4	2.0	1.0	0.2	1.0
5	10.0	1.0	1.7	2.2
6	10.0	2.0	3.0	1.3
	h	M	$a(M)$	$b(M)$

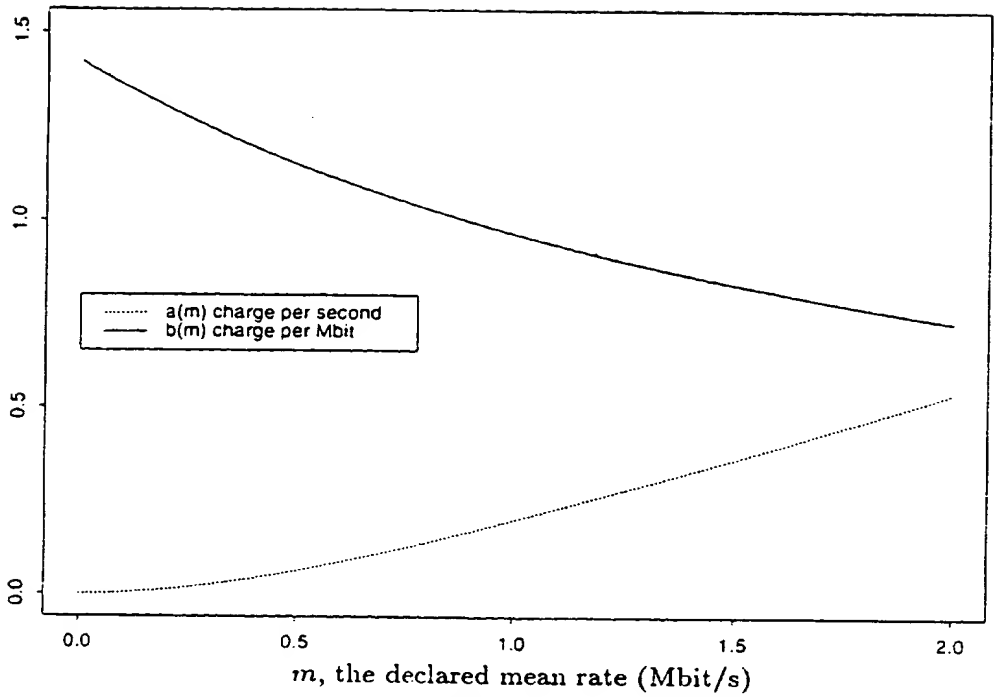


Figure 3. Dependence of tariff on declaration, for a peak rate of 2 Mbit/s.

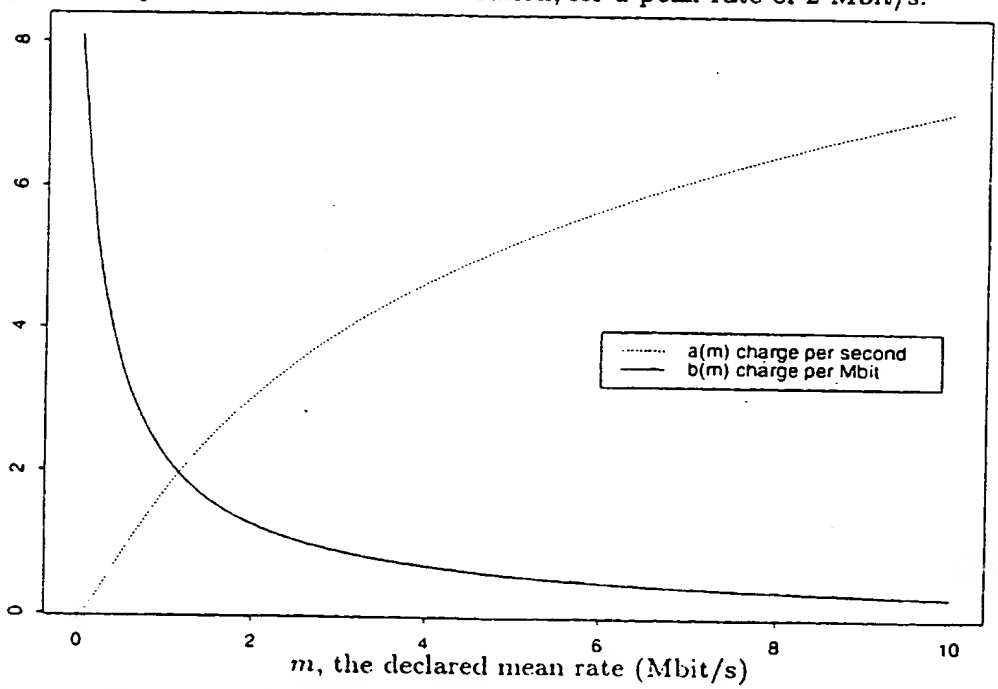


Figure 4. Dependence of tariff on declaration, for a peak rate of 10 Mbit/s.

Similarly tariffs may be calculated for sources with other peak rates. For a peak rate of 0.1 Mbit/s the bandwidth $B(M)$ is almost linear in M, producing a variable charge $b(m)$ per unit of traffic that is almost constant in m. Since statistical multiplexing is efficient for sources with such low peak rates, very little incentive need be given to determine mean rates accurately. Peak rates above 2 Mbit/s produce more concave effective bandwidths than that illustrated in Figure 1, and hence a more rapidly decaying charge $b(m)$ per unit of traffic, and a more rapidly increasing charge $a(m)$ per unit time: see Figure 4. It is interesting to note that for the three service types shown in Table 1 almost all of the total cost to the user arises from the variable charge, and that the variable charge is only slightly higher for service type 2 than 1, but much higher for service type 3. All three service types shown in Table 1 have a high peak-mean ratio. For service types with higher mean rates, such as those shown in Table 2, much more of the total cost arises from the fixed charge, more than half in the case of service type 6.

5. GENERALIZATIONS

We have obtained tariffs with desirable properties under the assumption that sources are on/off sources with known peak rate. The procedure generalizes straightforwardly to more general sources: the case of an on/off source with fixed but unknown peak rate is discussed in [10]. Indeed even the form (3) is not essential to the development: all that is necessary for our conclusions is that effective bandwidth be expressed as a concave function of the expectation of a measurable quantity. For further models where effective bandwidths possess this property see [4], [5].

We have concentrated attention on the case of a single resource, but aspects of our approach generalize to the case of multiple resources. For example suppose there are several possible routes through a network for a call, and let r label a route: write $k \in r$ if resource k lies on route r. Suppose that it costs the network c_k to provide an additional unit of effective bandwidth at resource k, and that sources are as described in Section 3, with known peak rate but uncertain mean rate. Then the natural measure of the cost per unit time of a call with mean rate M is

$$B(M) = \min_r \sum_{k \in r} c_k B_k(M).$$

where $B_k(M)$ is given by expression (4), with α replaced by α_k. Thus $B(M)$ is concave, since concavity is inherited under summation and minimization, and the associated tariff is just

$$\hat{f}(m; M) = \min_r \sum_{k \in r} c_k \hat{f}_k(m; M),$$

where $\hat{f}_k(m; M)$ is the tangent to the curve $B_k(M)$ at the point $M = m$. The optimal declaration under this tariff is just $\mathsf{E}_G M$, which is all the network needs to know to choose the route minimizing cost per unit time, and under this declaration the expected cost per unit time to the call and to the network are equal.

For expositional purposes we have phrased the development of this paper in terms of 'users' and a 'network'. We conclude by noting that this is just one interpretation: if users and the network form a single operation, then the tariffs we have described provide a natural mechanism for decentralized decision-making and control; on the other hand, the tariffs are cost-based rather than market-driven, and would form just one ingredient of a tariff structure in environments where there is competition between networks for users.

REFERENCES

1. Boyer, P.E., Guillemin, F.M., Servel, M.J. and Coudrese, J.-P. (1992) Spacing cells protects and enhances utilization of ATM network links. *IEEE Network* **6**, September 1992, 38–49.
2. Decina, M. and Trecordi, V. eds. (1992) Traffic management and congestion control for ATM networks. *IEEE Network* **6**, No.5, September 1992.
3. De Veciana, G. and Walrand, J. (1993) Effective bandwidths: call admission, traffic policing and filtering for ATM networks. Department of Electrical Engineering and Computer Science, Berkeley.
4. Elwalid, A.I. and Mitra, D. (1993) Effective bandwidth of general Markovian traffic sources and admission control of high speed networks. *IEEE–ACM Trans. Networking* **1**.
5. Gibbens, R.J. and Hunt, P.J. (1991) Effective bandwidths for the multi-type UAS channel. *Queueing Systems* **9**, 17–28.
6. Griffiths, T.R. (1990) Analysis of a connection acceptance strategy for asynchronous transfer mode networks. Globecom 90, paper 505.4.
7. Hui, J.Y. (1988) Resource allocation for broadband networks. *IEEE Selected Areas in Commun.* **6**, 1598–1608.
8. Hui, J.Y. (1990) *Switching and Traffic Theory for Integrated Broadband Networks.* Kluwer, Boston.
9. Kelly, F.P. (1991) Effective bandwidths at multi-class queues. *Queueing Systems* **9**, 5–16.
10. Kelly, F.P. (1994) On tariffs, policing and admission control for multiservice networks. *Operations Research Letters.*
11. Lindberger, K. (1991) Analytical methods for the traffical problems with statistical multiplexing in ATM-networks. ITC 13, Copenhagen. In *Teletraffic and Datatraffic* (ed. A. Jensen and V.B. Iverson), Elsevier.
12. Mitra, D. and Mitrani, I. (1991) Editorial introduction to Communication Systems. Special issue of *Queueing Systems* **9**, 1–4.
13. Roberts, J.W. ed. (1992) Performance evaluation and design of multiservice networks. Commission of the European Communities.
14. Trajković, L. and Golestani, S.J. (1992) Congestion control for multimedia services. *IEEE Network* **6**, September 1992, 20–26.
15. Whitt, W. (1992) Tail probabilities with statistical multiplexing and effective bandwidths in multi-class queues. A.T. & T. Bell Laboratories.